EUROPEAN ASSOCIATION FOR MACHINE TRANSLATION

EAMT2018

Proceedings of the

21st Annual Conference of the European Association for Machine Translation

28–30 May 2018

Universitat d'Alacant

Alacant, Spain

Edited by

Juan Antonio Pérez-Ortiz

Felipe Sánchez-Martínez

Miquel Esplà-Gomis

Maja Popović

Celia Rico

André Martins

Joachim Van den Bogaert

Mikel L. Forcada

Organised by

Universitat d'Alacant

Universidad de Alicante

transducens

research group

ISBN: 978-84-09-01901-4

Spelling Normalization of Historical Documents by Using a Machine Translation Approach

Miguel Domingo
PRHLT Research Center
Universitat Politècnica de València
midobal@prhlt.upv.es

Francisco Casacuberta
PRHLT Research Center
Universitat Politècnica de València
fcn@prhlt.upv.es

Abstract

The lack of a spelling convention in historical documents makes their orthography to change depending on the author and the time period in which each document was written. This represents a problem for the preservation of the cultural heritage, which strives to create a digital text version of a historical document. With the aim of solving this problem, we propose three approaches—based on statistical, neural and character-based machine translation—to adapt the document's spelling to modern standards. We tested these approaches in different scenarios, obtaining very encouraging results.

1 Introduction

With the aim of preserving the cultural heritage, there is an increased need for the digitalization of historical documents, a procedure which strives for creating digital text which can be searched and automatically processed (Piotrowski, 2012). However, the linguistic properties of historical documents create an additional difficulty. On the one hand, human language evolves with the passage of time. On the other hand, the lack of a spelling convention makes orthography to change depending on the author and the time period in which a given document was written. This makes historical documents harder to read, and makes it even more difficult to search for certain information in a collection of documents, or any other process that must be applied to them.

Spelling normalization aims to resolve these problems. Its goal is to adapt the document's spelling to modern standards, increasing documents' readability and achieving an orthography consistency. Some approaches to spelling normalization include creating an interactive tool that includes spell checking techniques to assist the user in detecting spelling variations (Baron and Rayson, 2008). Porta et al. (2013) made use of a weighted finite-state transducer, combined with a modern lexicon, a phonological transcriber and a set of rules. Scherrer and Erjavec (2013) combined a list of historical words, a list of modern words and character-based Statistical Machine Translation (SMT). Bollmann and Søgaard (2016) took a multi-task learning approach using a deep bi-LSTM applied at a character level. Ljubešic et al. (2016) applied a token/segment-level character-based SMT approach to normalize historical and user-created words. Domingo et al. (2017) applied a SMT approach combined with the use of data selection techniques. Finally, Korchagina (2017) made use of rule-based MT, character-based SMT and character-based NMT.

In this work, we propose three approaches to tackle spelling normalization: a method based on SMT; another method based on Neural Machine Translation (NMT); and another method based on Character-Based Machine Translation (CBMT). Our main contribution are the followings:

- First use (to the best of our knowledge) of word-based and subword-based NMT—character-based NMT was already used by Korchagina (2017)—for spelling normalization.
- Comparison of different approaches based on SMT and NMT.
- Experimented with four historical corpora

Pérez-Ortiz, Sánchez-Martínez, Esplà-Gomis, Popović, Rico, Martins, Van den Bogaert, Forcada (eds.)
Proceedings of the 21st Annual Conference of the European Association for Machine Translation, p. 129–137
Alacant, Spain, May 2018.

from three different time periods, in two different languages and with three distinct alphabets.

The rest of this document is structured as follows: In Section 2, we introduce the machine translation approaches used in our work. Section 3 presents the different approaches taken to achieve spelling normalization. Then, in Section 4, we describe the experiments conducted in order to assess our proposal. After that, in Section 5, we present and discuss the results of those experiments. Finally, in Section 6, conclusion are drawn.

2 Machine Translation Approaches

In this section, we present the machine translation approaches used in our work.

2.1 Statistical Machine Translation

The goal of SMT is to find, given a source sentence $\mathbf{x}$, its best translation $\hat{\mathbf{y}}$ (Brown et al., 1993):

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} Pr(\mathbf{y} \mid \mathbf{x}) \quad (1)$$

For years, phrase-based models (Koehn, 2010) have been the prevailing approach to compute this expression. These models rely on a log-linear combination of different models (Och and Ney, 2002): namely, phrase-based alignment models, reordering models and language models; among others (Zens et al., 2002; Koehn et al., 2003). However, more recently, this approach has shifted into neural models (see Section 2.2).

2.2 Neural Machine Translation

NMT is the neural approach to compute Eq. (1). Frequently, it relies on a Recurrent Neural Network (RNN) encoder-decoder framework. At the encoding step, the source sentence is projected into a distributed representation. Then, at the decoding step, the decoder generates its translation word by word (Sutskever et al., 2014).

The system's input is a sequence of words in the source language. Each source word is linearly projected to a fixed-sized real-valued vector through an embedding matrix. These word embeddings are feed into a bidirectional (Schuster and Paliwal, 1997) Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) network, resulting in a sequence of annotations produced by concatenating the hidden states from the forward and backward layers.

The model features an attention mechanism (Bahdanau et al., 2015), which allows the decoder to focus on parts of the input sequence, computing a weighted mean of annotations sequence. These weights are computed by a soft alignment model, which weights each annotation with the previous decoding state.

The decoder is another LSTM network, conditioned by the representation computed by the attention model and the last word generated. Finally, a deep output layer (Pascanu et al., 2013) computes a distribution over the target language vocabulary.

The model is trained by means of stochastic gradient descend, applied jointly to maximize the log-likelihood over a bilingual parallel corpus. At decoding time, the model approximates the most likely target sentence with beam-search (Sutskever et al., 2014).

2.3 Character-based Machine Translation

CBMT comes as a solution to reduce the training vocabulary by dividing words into a sequence of characters, and treating each character as if it were a word. Moreover, it also strikes for being a solution of not having a perfect segmentation algorithm—which should be able to segment a given sentence in any language, into a sequence of lexemes and morphemes (Chung et al., 2016).

Although CBMT was already being researched in SMT (Tiedemann, 2009; Nakov and Tiedemann, 2012), its interest has increased with NMT. Some approaches to character-based NMT consist in using hierarchical NMT (Ling et al., 2015), a character level decoder (Chung et al., 2016), a character level encoder (Costa-Jussà and Fonollosa, 2016) or, for alphabets in which words are composed by fewer characters, by constructing an NMT system that takes advantage of that alphabet (Costa-Jussà et al., 2017).

3 Spelling Normalization

In this section, we propose different approaches to adapt the spelling of historical documents to modern standards.

Our first approach is based on SMT. Considering the document's language as the source language and its normalized version of that language as the target language, we propose to use SMT to adapt the document's spelling to modern standards.

In our second approach, we wanted to assess how well NMT works for normalizing the spelling of a

historical document. Therefore, similarly as to with the previous approach, considering the document's language as the source language and its normalized version of that language as the target language, we propose to use NMT to adapt the document's spelling to modern standards.

Finally, since in spelling normalization changes frequently occur at a character level, it seemed fitting to use a character-based strategy. Therefore, our third approach is based on CBMT. Similarly as to with the previous approaches, considering the document's language as the source language and its normalized version of that language as the target language, we propose to use CBMT to adapt the document's spelling to modern standards.

As a starting point and to have the same conditions in both SMT and NMT, in this work we chose to use the simplest character-based approach: to split words into characters and, then, apply conventional SMT/NMT.

4 Experiments

In this section, we describe the experiments conducted in order to assess our proposal. Additionally, we present the corpora and metrics.

4.1 Corpora

To conduct our experiments, we made use of the following corpora:

Entremeses y Comedias (F. Jehle, 2001): A collection of comedies by Miguel de Cervantes, written in 17^{th} century Spanish.

Quijote (F. Jehle, 2001): The 17^{th} century Spanish novel by Miguel de Cervantes.

Bohorič (Ljubešić et al., 2016): A collection of 18^{th} century Slovene texts written in the Bohorič alphabet.

Gaj (Ljubešić et al., 2016): A collection of 19^{th} century Slovene texts written in the Gaj alphabet.

The first two corpora are Spanish literary works, written across the 17^{th} century. The first corpus is composed of 16 plays—8 of which have a very short length—while the second corpus is a two-volumes novel. The last two corpora are a collection of texts extracted from Slovene books. The first one is made up of texts from the 18^{th} century and it is written in the old Bohorič alphabet, and the second one is made up of texts from the 19^{th} century and written in the contemporary Gaj alphabet. Table 1 shows the corpora statistics.

4.2 Metrics

In order to asses our proposal, we made use of the following well-known metrics:

BiLingual Evaluation Understudy (BLEU) (Papineni et al., 2002): computes the geometric average of the modified n-gram precision, multiplied by a brevity factor that penalizes short sentences.

Translation Error Rate (TER) (Snover et al., 2006): computes the number of word edit operations (insertion, substitution, deletion and swapping), normalized by the number of words in the final translation.

Character Error Rate (CER): computes the number of character edit operations (insertion, substitution and deletion), normalized by the number of characters in the final translation.

Confidence intervals ($p = 0.05$) are computed for all metrics by means of bootstrap resampling (Koehn, 2004).

4.3 Systems

SMT systems were trained with the `Moses` toolkit (Koehn et al., 2007), following the standard procedure: we optimized the weights of the log-lineal model with MERT (Och, 2003), and used `SRILM` (Stolcke, 2002) to estimate a 5-gram language model, smoothed with the improved Kneser-Ney method (Chen and Goodman, 1996). Moreover, since source and target have the same linguistic structures—the only changes between source and target are orthographic—we used monotonous reordering. Finally, the corpora were lowercased and tokenized using the standard scripts, and the translated text was truecased with `Moses`' truecaser.

NMT systems were trained with `OpenNMT` (Klein et al., 2017), as described in Section 2.2. Following the findings from Britz et al. (2017), we used LSTM units. The size of the LSTM and word embedding were set according to the results of the development set. We used *Adam* (Kingma and Ba, 2014) with a learning rate of 0.0002 (Wu et al., 2016). The beam size was set to 6. Finally, the corpora were

		Entremeses y Comedias	Quijote	Bohorič	Gaj
Train	$\|S\|$	35.6K	48.0K	3.6K	13.0K
	$\|T\|$	250.0/244.0K	436.0/428.0K	61.2/61.0K	198.2/197.6K
	$\|V\|$	19.0/18.0K	24.4/23.3K	14.3/10.9K	34.5/30.7K
	$\|W\|$	52.4K	97.5K	33.0K	32.7K
Development	$\|S\|$	2.0K	2.0K	447	1.6K
	$\|T\|$	13.7/13.6K	19.0/18.0K	7.1/7.1K	25.7/25.6K
	$\|V\|$	3.0/3.0K	3.2/3.2K	2.9/2.5K	8.2/7.7K
	$\|W\|$	1.9K	4.5K	3.8K	4.5K
Test	$\|S\|$	2.0K	2.0K	448	1.6K
	$\|T\|$	15.0/13.3K	18.0/18.0K	7.3/7.3K	26.3/26.2K
	$\|V\|$	2.7/2.6K	3.2/3.2K	3.0/2.6K	8.4/8.0K
	$\|W\|$	3.3K	3.8K	3.8K	4.8K

Table 1: Corpora statistics. $|S|$ stands for number of sentences, $|T|$ for number of tokens, $|V|$ for size of the vocabulary and $|W|$ for the number of words whose spelling does not match modern standards. K denotes thousand.

lowercased and tokenized—and, later, truecased and detokenized—using `OpenNMT`'s tools.

CBMT systems were trained in the same way as conventional SMT/NMT systems. The only difference is that the corpora's words were previously split into characters. Then, after translating the document, words were restored.

To reduce the vocabulary, we used Byte Pair Encoding (BPE) (Sennrich et al., 2016). These systems were trained in the same way as conventional SMT/NMT systems. The only difference is that the corpora were previously encoded using BPE, and the translated text was decoded afterwards. BPE encoding was learned and applied using the scripts kindly provided by Sennrich et al. (2016). In learning the encoding, we used the default values for the number of symbols to create and the minimum frequency to create a new symbol.

Finally, in order to assess our proposal, we considered as a baseline the quality of the original document with respect to its ground truth version, in which the spelling has already been updated to match modern standards. Nonetheless, as a second baseline, we implemented a statistical dictionary. Using `mgiza` (Gao and Vogel, 2008), we computed *IBM's model 1* (Och and Ney, 2003) to obtain word alignments from source and target of the training set. Then, for each source word, we selected as its translation the target word which had the highest alignment probability with that source word. Finally, at translation time, we translated each source word with the translation that appeared in the dictionary. If a given word did not appear in the dictionary, then we left it untranslated.

5 Results

In this section, we present and discuss the experiments conducted in order to assess our proposal. Table 2 presents the experimental results.

The Slovene language had a big restructuring in the 18^{th} century. For this reason, *Bohorič*—whose documents were written during this period—is the corpus whose orthography differs the most compared to modern standards. Evaluating the document's spelling differences with respect to modern orthography results in a low BLEU value, a high TER value and a fairly high CER value. However, just by applying a statistical dictionary we achieved great improvements: BLEU and TER improved highly, and CER decreased significantly.

With our first approach, we achieved even greater improvements for all metrics. Furthermore, when using BPE to reduce the vocabulary, we achieved new improvements. These improvements were more notorious when evaluating with CER and BLEU, although they were significant with TER as well.

Our second approach achieved less satisfying results. The document's spelling differences were significantly reduced when measuring with BLEU and TER. However, the results were significantly worse than the ones obtained using a statistical dictionary. Furthermore, using CER to measure the spelling differences resulted in the document having more differences than the original document. Using BPE to reduce the vocabulary did not help. In fact, results were significantly worse. Most likely, this was due to the properties of the corpus: being the smallest of the corpora (less than four thousand sentences and with a vocabulary of around ten thou-

System	Entremeses y Comedias			Quijote			Bohorič			Gaj		
	BLEU	TER	CER	BLEU	TER	CER	BLEU	TER	CER	BLEU	TER	CER
Baseline	46.1 ± 1.4	31.7 ± 1.2	12.0 ± 0.4	59.6 ± 1.2	19.4 ± 0.7	7.4 ± 0.3	16.4 ± 1.6	49.0 ± 1.5	21.7 ± 0.6	68.1 ± 1.1	12.3 ± 0.5	3.5 ± 0.1
SD	80.8 ± 1.2	8.3 ± 0.5	4.0 ± 0.3	89.7 ± 0.8	5.3 ± 0.5	3.4 ± 0.3	52.5 ± 2.0	20.7 ± 1.2	17.2 ± 0.7	75.1 ± 0.8	8.8 ± 0.4	8.7 ± 0.3
SMT	82.1 ± 1.1	8.0 ± 0.5	6.7 ± 0.2	91.1 ± 0.7	4.5 ± 0.4	5.3 ± 0.3	63.0 ± 2.1	15.1 ± 1.1	9.0 ± 0.5	$\mathbf{82.6 \pm 0.7}$	5.2 ± 0.3	2.8 ± 0.1
SMT_{BPE}	83.6 ± 1.1	7.2 ± 0.5	6.2 ± 0.2	$\mathbf{94.6 \pm 0.6}$	$\mathbf{2.8 \pm 0.3}$	4.3 ± 0.2	70.4 ± 2.0	11.7 ± 1.0	5.3 ± 0.3	$\mathbf{83.7 \pm 0.7}$	$\mathbf{1.8 \pm 0.3}$	2.7 ± 0.1
NMT	72.2 ± 1.4	15.2 ± 0.9	18.0 ± 0.8	84.4 ± 0.9	8.1 ± 0.5	10.2 ± 2.4	36.7 ± 2.0	33.9 ± 2.1	41.4 ± 1.4	50.4 ± 1.4	28.3 ± 3.3	36.0 ± 2.7
NMT_{BPE}	76.7 ± 1.3	12.4 ± 0.8	8.1 ± 0.5	92.0 ± 0.7	4.6 ± 0.4	3.8 ± 0.3	31.6 ± 2.2	43.5 ± 6.1	48.6 ± 3.6	68.0 ± 1.5	23.7 ± 3.7	19.8 ± 2.6
CBSMT	$\mathbf{91.4 \pm 0.9}$	$\mathbf{3.7 \pm 0.4}$	$\mathbf{1.2 \pm 0.1}$	$\mathbf{94.7 \pm 0.6}$	$\mathbf{2.8 \pm 0.3}$	$\mathbf{2.0 \pm 0.2}$	$\mathbf{75.5 \pm 1.8}$	$\mathbf{8.7 \pm 0.9}$	$\mathbf{2.4 \pm 0.2}$	$\mathbf{83.2 \pm 0.7}$	5.0 ± 0.3	$\mathbf{1.3 \pm 0.1}$
CBNMT	81.3 ± 1.3	8.3 ± 0.8	3.0 ± 0.6	91.0 ± 0.7	4.6 ± 0.4	2.9 ± 0.3	27.6 ± 2.4	85.2 ± 6.7	68.2 ± 4.5	40.2 ± 1.9	62.7 ± 2.9	52.5 ± 2.1

Table 2: Experimental results. Baseline system corresponds to considering the original document as the document to which the spelling has been normalized to match modern standards. SD is the statistical dictionary. SMT is the standard SMT system. SMT_{BPE} is the SMT system trained after encoding the corpora using BPE. NMT is the standard NMT system. NMT_{BPE} is the NMT system trained after encoding the corpora using BPE. CBSMT is the character-based SMT system. CBNMT is the character-based NMT system. Best results are denoted in **bold**.

sand words), it was not big enough for NMT to learn properly how to update the document's orthography.

Finally, our third approach was both the most and least satisfying. While character-based SMT achieved the best results for all metrics–all of them were significantly better than the results achieved by SMT with BPE—character-based NMT achieved the worst results. Once more, this was most likely due to the corpus being too small for the neural systems.

Entremeses y Comedias, the oldest of the corpora, is the next corpus with greater orthographic difference. Nonetheless, the quality of the original document shows fairly good BLEU value, a considerable good TER value, and a low CER value. In spite of this, the statistical dictionary achieved significant improvements, the most noteworthy being the increase of BLEU, which was the metric that showed the lowest quality.

The SMT approach reduced significantly the spelling differences from the original document. However, in this case, results were not significantly different to the ones obtained by the statistical dictionary, except when evaluating with CER, which results were slightly (around two CER points) worse. Moreover, reducing the vocabulary with BPE did not achieved a significant difference with using the full vocabulary.

The NMT system behave in a similar fashion as with the previous corpus: BLEU and TER showed a significant reduction of the spelling difference from the original document, but smaller than the reduction achieved by the statistical dictionary. In this case, however, the differences with the statistical dictionary were smaller (around eight points of BLEU and TER). Moreover, although CER still showed more spelling differences than in the original document, its value was not as bad as with *Bohorič* (around six points of CER). Furthermore, despite still being worse than the statistical dictionary, BPE helped to improve results. It is worth noting the improvement in CER (around ten points), which represents an improvement with respect to the spelling differences in the original document.

Once more, the character-based approach yielded the best results. Character-based NMT was the neural approach which yielded the best results, although these results were not significantly different to the ones obtained by the statistical dictionary. However, character-based SMT did significantly improved the statistical dictionary.

Similarly to what happened with the other corpora, the statistical dictionary significantly reduced the spelling differences in the *Quijote* corpus. It is worth noting, however, that these differences are considerable smaller in this corpus: measuring the spelling differences in the original document shows a fairly good BLEU and TER values, and fairly small CER values.

In this case, the SMT approach did not yield results as satisfactorily as with the previous corpora. Results showed a significant improvement with respect to the original document. However, this improvement was not significantly different than the one achieved by the statistical dictionary—except when measuring with CER, whose value was significantly worse. Nonetheless, BPE improved the results, and the generated document was significantly better (for all metrics except for CER) than the document generated by the statistical dictionary.

The results yielded by the NMT system showed a significant improvement with respect to the spelling differences from the original document (except when measuring with CER), but this improvement was significantly worse than the one achieved by

the statistical dictionary. Reducing the vocabulary with BPE helped to improve the results—specially when measuring with CER, whose results were now significantly better than the original document—but they were similar to the statistical dictionary's results.

Finally, the character-based approached achieved, once more, the best results. However, while using CER to measure the document's spelling differences with respect to modern standards yielded a significant improvement (for character-based SMT), measuring with BLEU and TER yielded similar results to using the SMT approach combined with BPE. Similarly, character-based NMT achieved a significant improvement in terms of CER, but similar BLEU and TER results to the NMT-BPE approached.

Being the newest corpus, *Gaj* contains fewer spelling differences with respect to modern orthography. In fact, measuring the spelling differences from the original document already yielded satisfactory BLEU and TER values, and a low CER value. Nonetheless, the statistical dictionary managed to improve BLEU and TER results, although yielded a worse CER value.

The SMT approach managed to significantly improve results for all metrics. However, reducing the vocabulary with BPE yielded similar results, except when measuring with TER, whose results were significantly better.

Gaj being a fairly small corpus (thirteen thousand sentences and with a vocabulary of around thirty thousand words), the NMT systems behaved similarly as with *Bohorič*: The generated document had more spelling differences than the original document. Using BPE improved results, but the generated document still contained more spelling differences than the original one.

Character-based SMT yielded the best results when using CER to measure the spelling difference. However, measuring with BLEU and TER yielded similar results to the SMT approach. Character-based NMT, however, was the NMT approach which yielded the worst results—specially when measuring with TER and CER.

In general, except for one exception (*Gaj*, whose best results—when evaluating with TER—were achieved by the approach that combined SMT with BPE), character-based SMT was the approach that yielded the best results for all metrics. It is also worth noting how well—for being such a simplistic approach—using an statistical dictionary behave: except for one exception (*Gaj*, which yielded an increase of spelling differences when evaluating with CER), all results showed a significant reduction of spelling differences with respect to the original document and, in some cases, not too much worse than character-based SMT.

The BLEU and TER from the original document, and how much these values have significantly improved, seem to indicate that the final document is quite different to the original one. However, CER seems to indicate otherwise. Most likely, since spelling differences occur more frequently at a character level (i.e., most orthographic changes consist in a few letters per word), BLEU and TER—which evaluate at a word-level—are being penalized. Nonetheless, all metrics show that the spelling differences have been significantly reduced.

6 Conclusions and Future Work

In this work, we proposed three machine translation approaches to update the spelling of a historical document to match modern standards, increasing the document's readability and helping in the preservation of the cultural heritage.

Additionally, as an extra baseline, we proposed a simplistic approach: Based on the frequency of which, on the training corpora, the spelling of a word is changed, to build a statistical dictionary. Then, on a given document, we checked, word by word, if it was on the dictionary. If the search was positive, we changed that word by the translation that appeared in the dictionary. Otherwise, we left the word as it appeared in the original document.

We tested our proposal with four datasets formed by documents from three different time periods, two different languages and three distinct alphabets, obtaining very encouraging results.

In general, approaches based on SMT yielded better results than those based on NMT. This was specially true for the smallest corpora, in which the neural systems were not able to learn properly and yielded more spelling differences than the ones contained in the original document.

As it was to be expected due to the task characteristics (in spelling normalization, changes frequently occur at a character level), the character-based approaches—both phrased-based and neural—yielded the best results for each kind of system (i.e., character-based was the best SMT approach,

and character-based NMT was the best NMT approach). The exception was character-based NMT, which yielded worse results when applied to the smallest corpora.

Finally, it is worth noting how well the statistical dictionary behaves. Although its results were not the best, they were close enough to take this approach into consideration. Being the simplest and fastest to compute, it could be useful in cases in which its worth sacrificing quality to increase speed.

As a future work, we would like to try new character-based approaches. In this work, we tested the simplest approach (to split the words into characters and, then, apply conventional SMT/NMT). However, more complex approaches have been developed in recent years (Chung et al., 2016; Costa-Jussà and Fonollosa, 2016; Costa-Jussà et al., 2017).

Finally, a frequent problem when working with historical documents is the scarce availability of parallel training data (Bollmann and Søgaard, 2016). Therefore, we would like to obtain more diverse corpora to be able to experiment in broader domains: older time periods, documents written by a great variety of authors, etc. Additionally, we would like to explore the generation of synthetic data (Sennrich et al., 2015) to create new training data.

Acknowledgments

The research leading to these results has received funding from the Ministerio de Economía y Competitividad (MINECO) under project CoMUN-HaT (grant agreement TIN2015-70924-C2-1-R), and Generalitat Valenciana (grant agreement PROMETEO/2018/004). We gratefully acknowledge the support of NVIDIA Corporation with the donation of a GPU used for this research.

References

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (arXiv:1409.0473)*.

Baron, A. and Rayson, P. (2008). VARD2: A tool for dealing with spelling variation in historical corpora. *Postgraduate conference in corpus linguistics*.

Bollmann, M. and Søgaard, A. (2016). Improving historical spelling normalization with bi-directional lstms and multi-task learning. In *Proceedings of the International Conference on the Computational Linguistics*, pages 131–139.

Britz, D., Goldie, A., Luong, T., and Le, Q. (2017). Massive exploration of neural machine translation architectures. *arXiv preprint arXiv:1703.03906*.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Chen, S. F. and Goodman, J. (1996). An empirical study of smoothing techniques for language modeling. In *Proceedings of the Annual Meeting on Association for Computational Linguistics*, pages 310–318.

Chung, J., Cho, K., and Bengio, Y. (2016). A character-level decoder without explicit segmentation for neural machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1693–1703.

Costa-Jussà, M. R., Aldón, D., and Fonollosa, J. A. (2017). Chinese–spanish neural machine translation enhanced with character and word bitmap fonts. *Machine Translation*, 31:35–47.

Costa-Jussà, M. R. and Fonollosa, J. A. (2016). Character-based neural machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 357–361.

Domingo, M., Chinea-Rios, M., and Casacuberta, F. (2017). Historical documents modernization. *The Prague Bulletin of Mathematical Linguistics*, 108:295–306.

F. Jehle, F. (2001). *Works of Miguel de Cervantes in Old- and Modern-spelling*. Indiana University Purdue University Fort Wayne.

Gao, Q. and Vogel, S. (2008). Parallel implementations of word alignment tool. In *Proceedings of the Association for Computational Linguistics Software Engineering, Testing, and Quality Assurance Workshop*, pages 49–57.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. (2017). OpenNMT: Open-Source Toolkit for Neural Machine Translation. *arXiv preprint arXiv:1701.02810*.

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 388–395.

Koehn, P. (2010). *Statistical Machine Translation*. Cambridge University Press.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 177–180.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54.

Korchagina, N. (2017). Normalizing medieval german texts: from rules to deep learning. In *Proceedings of the Nordic Conference on Computational Linguistics Workshop on Processing Historical Language*, pages 12–17.

Ling, W., Trancoso, I., Dyer, C., and Black, A. W. (2015). Character-based neural machine translation. *arXiv preprint arXiv:1511.04586*.

Ljubešić, N., Zupan, K., Fišer, D., and Erjavec, T. (2016). Dataset of normalised slovene text KonvNormSl 1.0. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1068.

Ljubešic, N., Zupan, K., Fišer, D., and Erjavec, T. (2016). Normalising slovene data: historical texts vs. user-generated content. In *Proceedings of the Conference on Natural Language Processing*, pages 146–155.

Nakov, P. and Tiedemann, J. (2012). Combining word-level and character-level models for machine translation between closely-related languages. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 301–305.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 160–167.

Och, F. J. and Ney, H. (2002). Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 295–302.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistic*, 29(1):19–51.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Pascanu, R., Gulcehre, C., Cho, K., and Bengio, Y. (2013). How to construct deep recurrent neural networks. *arXiv preprint arXiv:1312.6026*.

Piotrowski, M. (2012). *Natural Language Processing for Historical Texts*. Number 17 in Synthesis Lectures on Human Language Technologies. Morgan & Claypool.

Porta, J., Sancho, J.-L., and Gómez, J. (2013). Edit transducers for spelling variation in old spanish. In *Proceedings of the workshop on computational historical linguistics*, pages 70–79.

Scherrer, Y. and Erjavec, T. (2013). Modernizing historical slovene words with character-based smt. In *Proceedings of the Biennial International Workshop on Balto-Slavic Natural Language Processing*, pages 58–62.

Schuster, M. and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Sennrich, R., Haddow, B., and Birch, A. (2015). Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of the Association for Machine Translation in the Americas*, pages 223–231.

Stolcke, A. (2002). SRILM - an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, pages 257–286.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks.

Tiedemann, J. (2009). Character-based PSMT for closely related languages. In *Proceedings of the Annual Conference of the European Association for Machine Translation*, pages 12–19.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Zens, R., Och, F. J., and Ney, H. (2002). Phrase-based statistical machine translation. In *Proceedings of the Annual German Conference on Advances in Artificial Intelligence*, volume 2479, pages 18–32.